*Czagány L. – Garai L. (szerk.) 2004: A szociális identitás, az információ és a piac. SZTE Gazdaságtudományi Kar Közleményei 2004. JATEPress, Szeged, 382-392. o.*

# Piaci információk és a multikollinearitás

*Petres Tibor*[1] *– Tóth László*[2]

*Nagy mennyiségű adatokból álló adatállományok nagyon gyakran kevés információt tartalmaznak. Ennek oka az adatállomány változói közötti kapcsolattal magyarázható. Ez a kapcsolat lényegében egyfajta redundanciaként is értelmezhető.*

*A tanulmányban egy új mérőszámot ismertetünk, amely a változók korrelációs mátrixának sajátértékeit tartalmazza, és lehetőséget nyújt a kollinearitás mértékének százalékos mérésére is: értéke 0 százalék, ha minden egyes sajátérték eggyel egyenlő és 100 százalék, ha az első kivételével az összes többi sajátérték nullával egyenlő.*

*Kulcsszavak: adatállomány, redundancia, multikollinearitás*

## 1. Kvantitatív elemzések

Az évezred elején, globalizálódó világunkban nagy mértékben növekszik mindannyiunk információigénye. Az adatok mennyiségének robbanásszerű növekedése nem jár együtt a megfelelő mértékű információ-növekedéssel. A két fogalom közötti jelentős különbséget az 1. ábra szemlélteti.

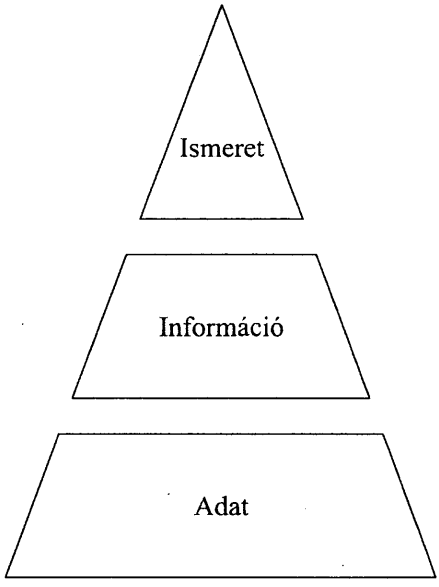
Igazából a döntéshozóknak nem az adatok hiányával, hanem azok bőségével kell szembenézniük, ugyanis (még a legóvatosabb becslések szerint is) az elektronikusan tárolt adatok volumene évente legalább megkétszereződik. A rendelkezésre álló adatok nagy mennyisége növeli ezen elemzésének összetettségét és az adatelemzőkkel szemben támasztott elvárásokat. Mivel az adatok információvá alakítása kisebb sebességgel történik, mint azok rendelkezésre bocsátása, a felhasználóknak egyre inkább adatelemzési szakértővé kell válniuk, ismerniük kell azokat a módszereket, amelyekkel az adatok értékelhetőek és hasznosíthatóak. Ebben segíthet a statisztika, mint a tömegjelenségek vizsgálatára szolgáló módszerek összessége.

A többváltozós statisztikai elemzéseknél két nézőpont ismeretes. Az egyik szerint az összes rendelkezésre álló változót szerepeltetjük, míg a másik szerint csak kevesebb változót használunk, amik azonban sűrítve tartalmazzák az (eredeti) adatállományban rejlő információt. Vagyis, képletesen szólva, az első szerint egy „narancs" egészét tekintjük, míg az utóbbi szerint ennek csak kivonatát, a „narancslét".

[1] Dr. Petres Tibor, egyetemi docens, SZTE Állam- és Jogtudományi Kar Statisztikai és Demográfiai Tanszék (Szeged)

[2] Tóth László, főosztályvezető, Informatikai és Hírközlési Minisztérium (Budapest)

*1. ábra* Adat-infotmáció-ismeret összefüggése




Rendszerezett információk összessége, problémák megoldását teszi lehetővé.

Döntéshozatalt szolgáló hasznos tartalmat hordozó adatok összessége. Minőségét az határozza meg, hogy milyen mértékben használható, alkalmazható.

Tárolt formájában független, tényszerű szám vagy szöveg. Minőségét pontossága, elérhetősége határozza meg.

## 2. Statisztikai modellek

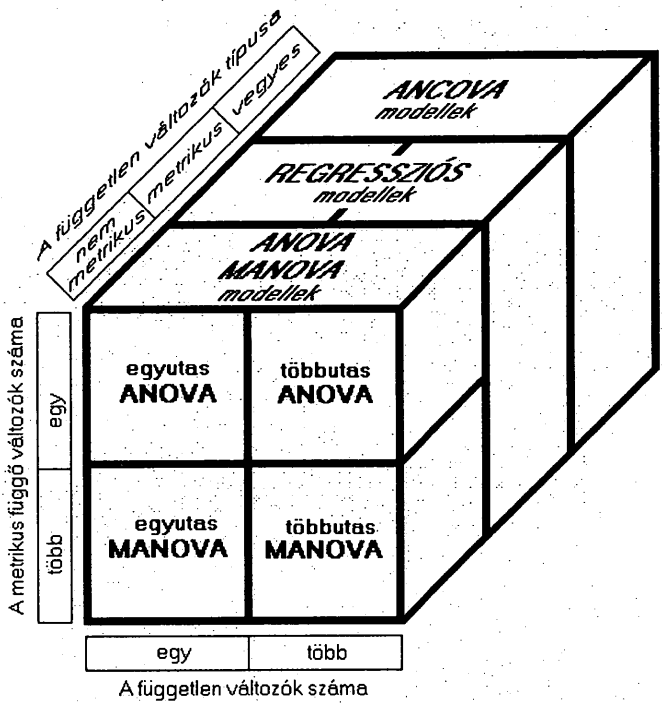
A fentiekből következően az alkalmazott modellek két csoportját lehet megkülönböztetni.

- Az ún. adatredukciós modellek esetén a változók számának csökkentésére törekszünk úgy, hogy ez a lehető legkevesebb információveszteséggel járjon. (Ebben az esetben nincs értelme a változók függő-független megkülönböztetésének.)
- Az ún. magyarázó modellek alkalmazásakor összefüggések feltárására törekszünk, vagy az összes rendelkezésre álló adat (illetve azokból képzett változó) alapján, vagy ezekből származtatott (kevesebb számú) változó(k) segítségével. Ebből következően megkülönböztetünk független (magyarázó-) és függő (eredmény-) változókat.

Az egyes magyarázó modellek alapvetően abban különböznek egymástól, hogy hány változóból állnak, illetve milyen mérési szintű adatokat tartalmaznak.

A legegyszerűbbek a kizárólag egy független- és egy függő változót tartalmazó modellek, leggyakrabban azonban több független és csak egy függő változónk van.

*2. ábra* A többváltozós statisztika modelljei




A függő változó szempontjából két nagy csoport létezik: az egyiknél a függő változó metrikus, míg a másiknál nemmetrikus. A független változók is lehetnek metrikus és nemmetrikus mérési szintűek, illetve egyszerre mindkét típusú változó szerepeltetése is előfordulhat.

A fentiek szerint a metrikus függő változó(ka)t tartalmazó modellek grafikus szemléltetése a 2. ábrán látható.

A 2. ábrán feltüntetett modellek túlnyomórészt lineáris összefüggések feltételezéséből indulnak ki, így ezek összefoglaló neve GLM (General Linear Model).

A 2. ábrán feltüntetett esetek közül a redundancia mérésének szempontjából kizárólag a metrikus adatok relevánsak. A metrikus adatok információtartalma az empirikus elemzéseknél lényeges kérdés, mert a nagyon nagy mennyiségű adat gyakran kevés információt hordoz, azaz nagyon nagy a redundancia mértéke. Redundancián a vizsgálat szempontjából újabb információt, érdemleges közlést már nem tartalmazó, „felesleges" adatokat értünk. Ennek a problematikának a bemutatása céljából a továbbiakban vizsgáljuk meg a regressziószámítást.

## 3. Regressziószámítás

Az egy eredmény- és több magyarázó változót tartalmazó regressziószámítás grafikus szemléltetése a 3. ábrán látható.

Amint látható, a bemeneti (ok) és a kimeneti (okozat) adatok összefüggése egyértelmű, azaz szerepük nem cserélhető fel. Az ezeket összekötő *f* funkcionális operátor egy fekete dobozként is felfogható. A regressziószámítás feladata ennek az operátornak az identifikálása.

### *3.1. A standard lineáris regressziós modell*

A többváltozós regressziós modell kompakt és kényelmesen kezelhető mátrixalgebrai jelölésmóddal:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

ahol **y** az eredményváltozó vektora, **X** a magyarázóváltozók mátrixa, $\boldsymbol{\beta}$ a regressziós paraméterek vektora, $\boldsymbol{\varepsilon}$ pedig a hibatagok vektora.

A modell specifikációjának fontos részét alkotják még az alábbiakban ismertetett feltételek is:

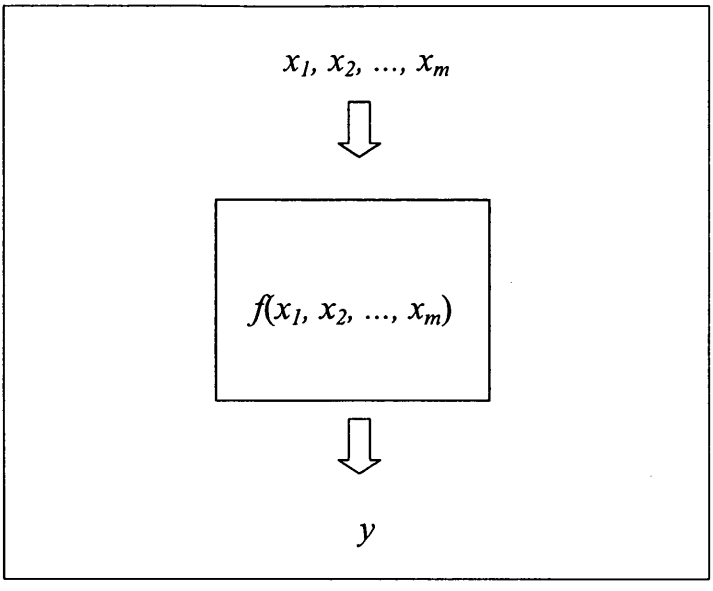
- A magyarázóváltozók nem sztochasztikusak (mérési hibát nem tartalmaznak), valamint lineárisan függetlenek (tehát nem redundánsak).
- A hibatagok nulla várható értékű, konstans varianciájú ($\sigma^2$), korrelálatlan valószínűségi változók, amik normális eloszlást követnek:

$$\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}).$$

A regressziószámítás gyakorlati alkalmazásakor ügyelnünk kell arra, hogy az említett modellt ne használjuk, ha valamelyik feltétele szignifikánsan nem teljesül! Közgazdasági elemzéseknél ennek leggyakrabban három oka lehet: multikollinearitás, autokorreláció, heteroszkedaszticitás. A továbbiakban részletesen az elsővel foglalkozunk, ami a redundanciának egyik megjelenési formájaként is felfogható.

A standard lineáris regressziós modell feltételezi, hogy a magyarázóváltozók egymástól lineárisan függetlenek. Ha valamelyik magyarázóváltozó kifejezhető a többi tényezőváltozó lineáris kombinációjaként, vagyis függvényszerű kapcsolatban áll a többi tényezőváltozóval, akkor teljes vagy extrém multikollinearitásról beszélünk. Ilyenkor az $\mathbf{X}'\mathbf{X}$ mátrix szinguláris, ezért nem invertálható. A teljes multikollinearitás felismerése könnyű, és egyszerűen megoldható az adott magyarázóváltozó elhagyásával. Az empirikus vizsgálatoknál azonban a magyarázóváltozók között inkább sztochasztikus kapcsolat jelentkezik.

*3. ábra* A regressziószámítás grafikus modellje




*3.2. A multikollinearitás következményei*

Ha a magyarázóváltozók egymástól lineárisan nem függetlenek, akkor az $n$ elemszámú mintán a legkisebb négyzetek módszerének közvetlen alkalmazásával kapott

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

becslés fontosabb tulajdonságai az alábbiak:

- A regressziós együtthatók standard hibái a

$$Var(\hat{\boldsymbol{\beta}}) = \frac{\mathbf{e}'\mathbf{e}}{n-m-1} \cdot (\mathbf{X}'\mathbf{X})^{-1} = s_e^2 \cdot (\mathbf{X}'\mathbf{X})^{-1}$$

  összefüggésből következően nőnek.
- Bizonytalanná, instabillá válnak a (továbbra is torzítatlan) becsléseink.
- Az egyes magyarázóváltozók hatásainak szeparált vizsgálata nem lehetséges, illetve a parciális regressziós együtthatók helyes értelmezése lehetetlenné válik.

A fentiek miatt a magyarázóváltozók kölcsönös függőségének mértékét mindig ellenőriznünk kell.

### *3.3. A multikollinearitás mérésének ismert mutatói*

A statisztikai szakirodalomban számos mutató ismert a multikollinearitás, illetve a redundancia számszerűsítésére. Most az alábbiakban ismertetünk néhányat a teljesség igénye nélkül.

Ha egy új magyarázóváltozót kapcsolunk be a modellbe, akkor a többszörös determinációs együttható vagy növekszik, vagy egyáltalán nem változik. Minden magyarázóváltozóra kiszámítva, hogy a modellbe utolsó változóként bevonva menynyivel növeli a determinációs együtthatót, ellenőrizhető a multikollinearitás. Ha az említett hatásoknak az összege egyenlő a többszörös determinációs együtthatóval, akkor azt mondhatjuk, hogy a magyarázóváltozók lineárisan függetlenek, azaz az adatok nem redundánsak. Ellenkező esetben az eredményváltozó szórásnégyzetének van olyan része, amit együttesen magyaráz több változó. A multikollineritás nagyságát ezzel az együttesen magyarázott résszel az alábbi módon mérhetjük.

$$M = r^2_{y\bullet x_1,x_2,\ldots,x_m} - \sum_{j=1}^{m}\left(r^2_{y\bullet x_1,x_2,\ldots,x_m} - r^2_{y\bullet x_1,x_2,\ldots,x_{j-1},x_{j+1},\ldots,x_m}\right)$$

Minél kisebb az eltérés az $M$ mutató értéke és a többszörös determinációs együttható között, annál jelentősebb a multikollinearitás, és ennek következtében a modell paramétereinek becslése mindinkább instabillá válik. Megjegyezzük továbbá, hogy az $M$ mutató negatív értéket is felvehet.

A multikollinearitás mérőszámaként a fenti logikának megfelelően használhatjuk az alábbi mutatót is. Az 1-nél nem nagyobb nemnegatív értékű

$$T_j = 1 - r^2_{j\bullet 1,2,\ldots,j-1,j+1,\ldots,m}$$

kifejezést tolerancia-mutatónak nevezzük. Ha a $j$-edik tényezőváltozó független a többi magyarázóváltozótól, akkor $T_j$ értéke 1. Ha $T_j = 0$, akkor extrém multikollinearitásról beszélünk.

A $T_j$ mutató reciprokát $VIF_j$-vel jelöljük (a Variance Inflation Factor rövidítése). Ez megmutatja, hogy a multikollinearitás, azaz az adatállomány redundanciája miatt milyen mértékben növekszik a becsült paraméterek varianciája, de magáról a redundancia mértékéről keveset mond.

Egy másik megközelítést használ a BARTLETT-féle próba, amely azt vizsgálja, hogy a változóink korrelációs mátrixa mennyire hasonlít egy egységmátrixhoz, vagyis változóink páronként korrelálatlanok-e. A teszt egy $\chi^2$-próbán alapul, aminek nullhipotézise a korrelációs mátrix és az egységmátrix egyezősége.

Az eddigiektől eltérően a magyarázóváltozók egészére vonatkozóan is ismert egy a multikollinearitást mérő mutatószám, amely Belsley (1980) nyomán a következő:

$$\gamma = \sqrt{\frac{\lambda_{max}}{\lambda_{min}}},$$

ahol $\lambda_{max}$ és $\lambda_{min}$ a normalizált magyarázóváltozók $\hat{\mathbf{X}}'\hat{\mathbf{X}}$ mátrixának legnagyobb illetve legkisebb sajátértéke. A $\gamma$ mutató értéke a magyarázóváltozók ortogonalitása, azaz a redundancia hiánya esetén 1. Több szerző szerint a mutató 30 feletti értéke utal erős multikollinearitásra, bár ez esetről esetre változhat.

Az ismertetett mutatók mindegyikének hátránya az, hogy az értelmezésük szubjektív és meglehetősen ellentmondásos. Az alábbiakban bemutatott eljárás előnye az, hogy alkalmazásával objektíven értelmezhető a multikollinearitás, azaz a redundancia mértéke.

## 4. A redundancia (multikollinearitás) új megközelítésű mérése

Az általunk javasolt mutató, sok más multikollinearitás-mutatóhoz hasonlóan, a magyarázóváltozók **R** korrelációs mátrixából indul ki. Az első lépésben kiszámítjuk az $m$ dimenziós **R** sajátértékeit ($\lambda_j$, $j$=1,2,...,$m$). Mivel a korrelációs mátrix pozitív szemidefinit, azaz a sajátértékek nem negatív számok, ezért kiszámíthatjuk szóródásuk mértékét, például a $v_\lambda$ relatív szórással

$$v_\lambda = \frac{\sigma_\lambda}{\bar{\lambda}} = \frac{\sqrt{\frac{\sum_{j=1}^{m}(\lambda_j - \bar{\lambda})^2}{m}}}{\frac{\sum_{j=1}^{m}\lambda_j}{m}} = \sqrt{\frac{\sum_{j=1}^{m}(\lambda_j - 1)^2}{m}} = \sigma_\lambda,$$

ami a

$$\sum_{j=1}^{m}\lambda_j = m$$

összefüggés miatt $\sigma_\lambda$ szórással egyenlő. Ennek a két mutatónak $\lambda_j \geq 0$ esetén a legnagyobb értéke $\sqrt{m-1}$.

Ezt a tulajdonságát felhasználva a mindenkori kapott eredményt normáljuk a

$$Red = \frac{\sigma_\lambda}{\sqrt{m-1}}$$

összefüggés szerint.

*1. táblázat* Néhány élelmiszer fogyasztásával kapcsolatos adat

| Évek | Sörfogy. | Egy főre jutó reáljöv. | Borfogy. | Gyümölcs fogy. | Sörár | Borár | Colaár | Pálinkaár | Fogyárindex |
|---|---|---|---|---|---|---|---|---|---|
| | liter/fő | % | liter/fő | kg | 0,5 liter | liter | 2 liter | 0,2 liter | % |
| 1990 | 105,3 | 100,9 | 27,7 | 72,3 | 16,5 | 60,5 | 55,2 | 62,8 | 100,0 |
| 1991 | 100,6 | 99,2 | 28,9 | 70,6 | 20,2 | 69,3 | 73,2 | 78,7 | 135,0 |
| 1992 | 94,0 | 95,7 | 29,8 | 72,8 | 22,8 | 70,4 | 82,2 | 91,6 | 166,1 |
| 1993 | 82,9 | 91,1 | 31,5 | 76,7 | 30,9 | 73,3 | 94,4 | 111,0 | 203,4 |
| 1994 | 84,7 | 93,5 | 29,2 | 70,2 | 34,0 | 83,0 | 119,0 | 124,0 | 241,6 |
| 1995 | 75,3 | 88,4 | 26,6 | 58,3 | 44,9 | 110,0 | 132,6 | 166,0 | 309,7 |
| 1996 | 71,3 | 87,8 | 30,3 | 64,4 | 55,6 | 152,0 | 165,4 | 208,0 | 382,8 |
| 1997 | 69,5 | 88,6 | 31,9 | 62,6 | 66,3 | 176,0 | 153,0 | 254,0 | 452,9 |
| 1998 | 69,3 | 91,8 | 33,6 | 68,5 | 73,9 | 188,0 | 169,0 | 285,0 | 517,7 |
| 1999 | 68,0 | 92,5 | 30,2 | 71,6 | 81,9 | 195,0 | 186,0 | 304,0 | 569,5 |
| 2000 | 71,6 | 96,5 | 28,3 | 110,8 | 89,4 | 209,0 | 197,0 | 317,0 | 625,3 |
| 2001 | 71,0 | 100,0 | 35,1 | 100,0 | 99,0 | 245,0 | 208,0 | 335,0 | 682,8 |

*Forrás*: KSH.

*2. táblázat* A fogyasztás 2001-es áron kifejezett adatai

| Évek | Sörfogy. | Egy főre jutó reáljöv. | Borfogy. | Gyümölcs fogy. | Sörár | Borár | Colaár | Pálinkaár |
|---|---|---|---|---|---|---|---|---|
| | liter/fő | % | liter/fő | kg | 0,5 liter | liter | 2 liter | 0,2 liter |
| 1990 | 105,3 | 100,9 | 27,7 | 72,3 | 112,7 | 413,1 | 376,9 | 428,8 |
| 1991 | 100,6 | 99,2 | 28,9 | 70,6 | 102,2 | 350,5 | 370,2 | 398,0 |
| 1992 | 94,0 | 95,7 | 29,8 | 72,8 | 93,8 | 289,5 | 338,0 | 376,7 |
| 1993 | 82,9 | 91,1 | 31,5 | 76,7 | 103,7 | 246,0 | 316,9 | 372,6 |
| 1994 | 84,7 | 93,5 | 29,2 | 70,2 | 96,1 | 234,6 | 336,3 | 350,4 |
| 1995 | 75,3 | 88,4 | 26,6 | 58,3 | 99,0 | 242,5 | 292,3 | 366,0 |
| 1996 | 71,3 | 87,8 | 30,3 | 64,4 | 99,2 | 271,1 | 295,0 | 371,0 |
| 1997 | 69,5 | 88,6 | 31,9 | 62,6 | 100,0 | 265,3 | 230,7 | 382,9 |
| 1998 | 69,3 | 91,8 | 33,6 | 68,5 | 97,5 | 248,0 | 222,9 | 375,9 |
| 1999 | 68,0 | 92,5 | 30,2 | 71,6 | 98,2 | 233,8 | 223,0 | 364,5 |
| 2000 | 71,6 | 96,5 | 28,3 | 110,8 | 97,6 | 228,2 | 215,1 | 346,1 |
| 2001 | 71,0 | 100,0 | 35,1 | 100,0 | 99,0 | 245,0 | 208,0 | 335,0 |

A redundancia hiánya, vagyis ortogonális magyarázóváltozók esetén a fenti mutató értéke 0, illetve 0 százalék, míg maximális redundancia (extrém multikollinearitás) esetén 1, illetve 100 százalék.

Így a *Red* mutató segítségével különböző adatállományok redundanciájának mértékét tudjuk számszerűsíteni. Mivel százalékban is kifejezhető mutatóról van szó, ezért különböző rangú *R* mátrixok fenti módon kiszámított mutatói közvetlenül összehasonlíthatóak, összevethetőek.

A *Red* mutató gyakorlati alkalmazásának lehetőségét legjobban egy példa segítségével szemléltethetjük. A kiinduló adatainkat az 1. táblázat tartalmazza.

Az 1. táblázat utolsó oszlopában szereplő fogyasztói árindexek segítségével a közölt termékek folyóáras egységárait deflálnunk kell. A 2001-es áron kifejezett mutatókat a 2. táblázat tartalmazza.

Ha az egy főre jutó sörfogyasztást befolyásoló (táblázatban feltüntetett) magyarázóváltozók információtartalmára vagyunk kíváncsiak, akkor a *Red* mutató segítségével számszerűsíteni tudjuk, hogy a sok adat milyen mértékben tartalmaz érdemleges közlést. A mutató kiszámításához szükségünk van a korrelációs mátrixra, ami a 3. táblázatban található.

Ennek a mátrixnak a sajátértékei az alábbiak:

$$\lambda_1 = 3{,}5817$$
$$\lambda_2 = 1{,}7130$$
$$\lambda_3 = 0{,}8830$$
$$\lambda_4 = 0{,}4838$$
$$\lambda_5 = 0{,}2239$$
$$\lambda_6 = 0{,}0841$$
$$\lambda_7 = 0{,}0306$$

Ezek összege a magyarázóváltozók számával egyenlő. Következő lépésben ki kell számítanunk a közölt sajátértékek szórását:

$$\sigma_\lambda = \sqrt{\frac{\sum_{j=1}^{7}(\lambda_j - 1)^2}{7}} = 1{,}1853.$$

Innen:

$$Red = \frac{1{,}1853}{\sqrt{7-1}} = 0{,}4839.$$

Ha a magyarázóváltozókat tartalmazó adatok egymástól lineárisan függetlenek, akkor a *Red* mutató értéke 0, azaz a redundancia mértéke 0 százalékos. A konk-

*3. táblázat* A korrelációs mátrix elemei

| | **Egy főre jutó reál-jövede-lem** | **Borfogy** | **Gyümölc sfogy** | **Sörár** | **Borár** | **Colaár** | **Pálinka-ár** |
|---|---|---|---|---|---|---|---|
| Egy főre jutó reáljövedelem ($x_1$) | 1,0000 | -0,0037 | 0,5627 | 0,3569 | 0,5504 | 0,2619 | 0,1946 |
| Borfogy ($x_2$) | -0,0037 | 1,0000 | 0,2211 | -0,2229 | -0,3303 | -0,5570 | -0,3541 |
| Gyümölcsfogy ($x_3$) | 0,5627 | 0,2211 | 1,0000 | -0,0830 | -0,2142 | -0,4059 | -0,4960 |
| Sörár ($x_4$) | 0,3569 | -0,2229 | -0,0830 | 1,0000 | 0,7488 | 0,4327 | 0,7287 |
| Borár ($x_5$) | 0,5504 | -0,3303 | -0,2142 | 0,7488 | 1,0000 | 0,6986 | 0,8875 |
| Colaár ($x_6$) | 0,2619 | -0,5570 | -0,4059 | 0,4327 | 0,6986 | 1,0000 | 0,6343 |
| Pálinkaár ($x_7$) | 0,1946 | -0,3541 | -0,4960 | 0,7287 | 0,8875 | 0,6343 | 1,0000 |

*4. táblázat* Az egyes változókhoz tartozó tolerancia-mutatók

| **Magyarázóváltozók** | $T_j$ | $VIF_j$ |
|---|---|---|
| Egy főre jutó reáljövedelem ($x_1$) | 0,14 | 6,90 |
| Borfogy ($x_2$) | 0,62 | 1,63 |
| Gyümölcsfogy ($x_3$) | 0,19 | 5,19 |
| Sörár ($x_4$) | 0,35 | 2,90 |
| Borár ($x_5$) | 0,05 | 19,95 |
| Colaár ($x_6$) | 0,30 | 3,39 |
| Pálinkaár ($x_7$) | 0,08 | 13,19 |

rét példában azonban a redundancia, azaz új információt, illetve érdemleges közlést már nem tartalmazó adatok mértéke 48,39 százalékos.

Összehasonlítás végett közöljük a többi ismertetett mutatónak a példa adatain felvett értékeit.

$$M = 0{,}9902 - \left((0{,}9902 - 0{,}9615) + \ldots + (0{,}9902 - 0{,}9729)\right) = 0{,}7706$$

Az egyes változókhoz tartozó tolerancia-mutatók értékei a 4. táblázat tartalmazza.

A magyarázóváltozók közül a $T_j = 0{,}05$ alapján arra lehet következtetni, hogy a borárakból képzett változó tartalmazza a legkevesebb érdemi többletinformációt.

A BARTLETT-féle próbához tartozó próbafüggvény értéke:

$$\chi^2 = 50{,}892;$$

ami alapján $v = 21$ szabadságfok mellett és 0,5 százalékos szignifikancia szinten a nullhipotézis elvetését jelenti, vagyis szignifikáns különbség van a korrelációs mátrix és a magyarázóváltozók ortogonalitását feltételező egységmátrix között.

A BELSLY-féle mutató értéke az alábbi:

$$\gamma = \sqrt{\frac{6{,}8963}{0{,}0004}} = 137{,}6.$$

Ennek 30-nál jóval nagyobb értéke szintén jelentős multikollinearitásra utal.

## 5. Összefoglalás

Empirikus elemzéseknél fontos tudni, hogy a nagymennyiségű adatot tartalmazó adatállományban mekkora a redundancia, azaz a sok adat milyen mértékben tartalmaz érdemleges közlést. A témára a regressziószámítás segítségével mutattunk rá, ahol ez a multikollinearitásból származó problémaként jelenik meg. Ennek mérésére a szakirodalomban többféle mutató ismert, de mindegyikre az jellemző, hogy értelmezésük szubjektív és meglehetősen ellentmondásos. A redundancia általunk bemutatott új megközelítésű mérése biztosítja a redundancia olyan számszerűsítését, amely (mivel normált és százalékban kifejezhető) egyértelműen értelmezhető. Ráadásul különböző adatállományokban mérni lehet az érdemleges közlés mennyiségét és azok mértékét is össze lehet közvetlenül hasonlítani.

*Felhasznált irodalom*

BELSLEY, D. - E. KUH - R. WELSCH 1980: Regression Diagnostics: Indentifying Influential Data and Sources of Collinearity, Wiley, New York.